Biosurveillance for Invasive Fungal Infections via Text Mining

David Martinez[1], Hanna Suominen[2], Michelle Ananda-Rajah[3], Lawrence Cavedon[1]

[1] NICTA, National ICT Australia and The University of Melbourne, Level 2 / Building 193, 3010 Melbourne, VIC, Australia
[2] NICTA, National ICT Australia and The Australian National University, Locked Bag 8001, 2601 Canberra, ACT, Australia
[3] Alfred Health and The University of Melbourne, Level 2, Burnet Institute, 85 Commercial Rd., 3004 Melbourne, VIC, Australia
david.martinez@nicta.com.au, hanna.suominen@nicta.com.au, m.ananda-rajah@alfred.org.au, lawrence.cavedon@nicta.com.au

Abstract. Invasive fungal diseases (IFDs) cause more than 1,000 deaths in hospitals and cost the health system more than AUD100m in Australia each year. The most common life-threatening IFD is aspergillosis and a patient with this IFD typically has 12 days prolonged in-patient time in hospital and an 8% mortality rate. Surveillance and detection of IFDs irrespective of the stage of diagnosis (i.e., early or late in disease) is important.

We describe an application of text mining techniques, using machine learning over a range of features, to automatically detect cases of patients with IFD from the text in the reports of CT scans performed on them. We focus on detecting the presence of aspergillosis; however, we anticipate the approach to be transferable to other diseases or conditions by training the text mining component over appropriate reports. Previous systems based on language technology have been deployed for processing radiology reports and for detecting hospital-acquired infection using language-processing technology, with significant success. Our approach differs by using a purely statistical/machine-learning approach to the language technology, and by being trained and tested on data collected from a number of hospitals.

We collected reports for 288 IFD and 291control patients from three different hospitals in Melbourne, Australia: Alfred Health, Melbourne Health, and Peter MacCallum Cancer Centre. We extracted a sample of 69 IFD and 49 control patients to perform detailed analysis of the text with regard to IFD; each patient had possibly multiple scans (and associated reports), resulting in a total of 398 scan reports from IFD-positive patients and 83 scan reports from control patients. We had medical experts annotate the patient-level classification on all scan reports at both sentence and report level: The annotators had to decide, for each sentence and report, whether it was positive, neutral, or negative with regards to IFD. We classify reports and patients as IFD-positive if they contain at least one positive sentence, and as negative otherwise.

We used the Weka SVM implementation and employed a variety of text- and concept-based features, including bag-of-words, punctuation, UMLS concepts and negated contexts extracted using MetaMap. We also automatically extract-

ed high-value terms (as measured using log-likelihood ratio) and formulated multi-word concept descriptions. Our system showed Sensitivity of 0.94 and Specificity of 0.76 for classifying individual reports as being indicative of aspergillus, and 1.0 and 0.51 for classifying patients as having contracted the infection.

Keywords: Biosurveillance; Clinical Reports; Machine Learning; Text Mining

1 Introduction

Invasive fungal diseases (IFDs) cause more than 1,000 deaths in hospitals and cost the health system more than AUD100m in Australia each year.[1] The most common life-threatening IFD is aspergillosis and a patient with this IFD typically has 12 days prolonged inpatient time in hospital and an 8% mortality rate.[1] Surveillance and detection of IFDs irrespective of the stage of diagnosis (i.e., early or late in disease) is important.

In this paper, we describe an application of text mining techniques, using machine learning (ML) over a range of features, to automatically detect cases of patients with IFD from the text in the reports of CT scans performed on them. In the description below, we focus on detecting the presence of *aspergillosis*; however, we anticipate the approach to be transferable to other diseases or conditions by training the text mining component over appropriate reports. Previous systems based on language technology have been deployed for processing radiology reports and for detecting hospital acquired infection using language-processing technology, with significant success.[2, 3] Our approach differs by using a purely statistical/machine-learning approach to the language technology, and by being trained and tested on data collected from a number of hospitals.

Our text mining technique will form the core of a notification and surveillance system that raises an alarm with a clinical team or nursing station overseeing the health of hospitalised patients. The resulting system can include a multitude of text classifiers, each trained to detect specified conditions over a written report for any scan performed on any patient. Ultimately, the text mining component will be part of a pervasive surveillance system which also monitors other types of data, such as lab results and images, and combines all pertinent information to produce high-accuracy detection. In addition to surveillance at hospital and patient levels, the system enables capturing and visualising the underlying evidence for a given classification decision at report and sentence levels.

2 Materials and Methods

Data collection: We collected data for patients with 288 IFD and 291 control patients from three different hospitals in Melbourne, Australia: Alfred Health, Melbourne Health, and Peter MacCallum Cancer Centre. The data consisted of the reports for all the scans performed on the patient over the hospitalisation period. For each scan, we use the written report (i.e., the radiologist's narrative) that describes the observations, and the state of the patient. From our initial set, we extracted a sample of 69 IFD and

49 control patients to perform detailed analysis of the text with regard to IFD; each patient had possibly multiple scans (and associated reports), resulting in a total of 398 scan reports from IFD-positive patients and 83 scan reports from control patients.

Annotation: We extended the patient-level classification by having medical experts annotate all the scan reports at both sentence and report level. The annotators had to decide, for each sentence and report, whether it was positive, neutral, or negative with regards to IFD. We used three assessors for the task, and performed double-blind annotation over the IFD patients in order to measure the annotation agreement. The process was performed as follows: the whole dataset was annotated by the main annotator, and the other two experts split the data, and each annotated half of the collection. The guidelines were initially set by the main annotator, and refined through discussion after annotating a small sample of the data. For control patients, we considered that the presence of IFD-positive annotations would be minimal, and we relied only on our main annotator for this task.

Methods: Our prediction system is based on supervised sentence classification, where a ML method predicts if a given sentence is indicative of IFD or not. We classify reports and patients as IFD-positive if they contain at least one positive sentence, and as negative otherwise. The underlying ML classifier is an implementation of Support Vector Machines (SVM) — we used the Weka toolkit's SVM implementation.[4] We employed a variety of text- and concept-based features, including bag-of-words, punctuation, UMLS concepts and negated contexts extracted using MetaMap.[5] We also automatically extracted high-value terms (as measured using the log-likelihood ratio test) and formulated multi-word concept descriptions (further details will be presented in full paper). For evaluation we used 10-fold cross-validation.

3 Results

Performance of the system at the report and patient levels (i.e., for detecting scans that seem to indicate evidence of aspergillosis and for classifying patients who have at least one (detected) positive report against the set of patients known to have been IFD-positive, respectively) is summarised in Table 1. High performance at report level is a prerequisite for any real-time detection system. Current work is being performed to improve on these results, but the values reported below are already quite strong. By comparison, the expert annotators' agreement over IFD-positive and IFD-negative sentences was 0.64 and 0.58 respectively, using Cohen's kappa.

Table 1. Performance at the report and patient levels. PPV = positive predictive value,

Level	Sensitivity	Specificity	Positive PV	Negative PV
Report	0.94	0.76	0.83	0.91
Patient	1.00	0.51	0.73	1.00

4 Conclusion

Future hospital information systems with IFD-surveillance capabilities would present an improvement on current practice where prospective IFD surveillance is not performed in the vast majority of centres. This surveillance requires high performance, especially sensitivity, at the patient level. Our current results suggest that this is feasible. Questions remain regarding installation in a hospital environment, including how to ensure system performance continues to improve if/when misclassifications are made. Focusing on surveillance (as opposed to real-time detection) simplifies this somewhat; we intend to explore this further in future planned trials.

Acknowledgements

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

1. Slavin M., J. Fastenau, I. Sukarm, P. Mavros, S. Crowley, W.C. Gerth, Burden of hospitalization of patients with *Candida* and *Aspergillus* infections in Australia, *International Journal of InfectiousDiseases*, 2004:8:111—120.
2. Friedman C, Alderson P, Austin J, Cimino JJ, and Johnson SB. A general natural language text processor for clinical radiology. *J. American Medical Informatics Association*, March 1994, 1(2):161--174.
3. Chapman WW, Fiszman M, Christensen L, Haug PJ. A comparison of classification algorithms to automatically identify chest x-ray reports that support pneumonia. *Journal of Biomedical Informatics*, 2001 Feb;34(1):4-1.
4. I. H.Witten and E. Frank, Data Mining: Practical machine learning tools and techniques. San Francisco, USA: Morgan Kaufmann, 2005.
5. A. R. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program," *AMIA Annual Symposium Proceedings*, (Washington DC), 17–21, 2001.